Detecting correlations among functional-sequence motifs

Detecting Correlations among Functional Sequence Motifs

Davide Pirino,[1] Jacopo Rigosa,[2] Alice Ledda,[3] and Luca Ferretti[4]

[1] *LEM, Scuola Superiore Sant'Anna, 56127 Pisa, Italy**

[2] *The BioRobotics Institute, Scuola Superiore Sant'Anna, 56127 Pisa, Italy*

[3] *Inserm U722, Faculté de Médecine Xavier Bichat, 75018 Paris, France*

[4] *Centre de Recerca en AgriGenòmica, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain*

Sequence motifs are words of nucleotides in DNA with biological functions, e.g. gene regulation. Identification of such words proceeds through rejection of Markov models on the expected motif frequency along the genome. Additional biological information can be extracted from the correlation structure among patterns of motif occurrences. In this paper a log-linear multivariate intensity Poisson model is estimated via expectation maximization on a set of motifs along the genome of *E. coli* K12. The proposed approach allows for excitatory as well as inhibitory interactions among motifs and between motifs and other genomic features like gene occurrences. Our findings confirm previous stylized facts about such types of interactions and shed new light on genome-maintenance functions of some particular motifs. We expect these methods to be applicable to a wider set of genomic features.

I. INTRODUCTION

Counting processes are the most natural way to model the occurrence of a particular type of event. Such a process is fully described by the probability P_t to observe the event a time t. Eventually, an additional variable S_t may indicate the actual state of the system [1] . The description of many physical, biological and social systems lies in the class of point processes in which the probability P_t^i of the i-th process is determined by the past history of all the processes that enter in the system, including the process itself (point processes with stochastic intensity). Related researches encompass very different fields such as photon counting, laser physics, astrophysics, geophysics, social phenomena and, as discussed in detail in this paper, genomics. For example, a self-exciting point process (usually called Hawkes' process, see [2]) is used in [3] to model the photomultiplier tubes' dark pulses: in this model an occurrence a time t_i of a dark pulse event increases the probability to observe another dark pulse for $t > t_i$, with an exponential decay interaction. An identical process is adopted by [4] to model a feedback-controlled cavity in a steady-state. On the same line, the occurrence of a photon count can be used to inhibit the probability of another photon count. Such a photon anticorrelation mechanism is used in [5–7] for the production of a particular state of light, namely photon-number-squeezed light. In astrophysics Hawkes' processes are introduced to model hotspots' interactions in accretion disc (see [8, 9]). A similar idea is behind the modeling of the small earthquake shocks that follow a main shock. Usually, in these models, several features are included for explaining the total amount of the observed intensity. In [10] the seismic activity (i.e. the probability of a seismic event) is described by a point process with stochastic intensity that includes self-excitation as well as trends, periodicity and interactions with other earthquakes in other locations (mutually exciting point processes). Not only the simple occurrence of a shock can increase the probability of subsequent shocks, but also its magnitude, as encompassed by the Epidemic-Type Aftershock Sequence (ETAS) model, which is the baseline model used in [11–13]. Finally, several kinds of social phenomena originated by complex network interactions are successfully described by point processes with stochastic intensity (as an example, think about the spatio-temporal pattern of a disease or opinion spreading). Among many we suggest the analysis of the book sale dynamics proposed by [14], in which the probability of a buy is conditioned by all the previous buys, as in an epidemic or avalanche model.

Given the recognized ductility of stochastic intensity point processes, in this paper we propose the adoption of such a process for the detection of statistical interactions among events along a string of DNA. The type of events we have in mind can be either the occurrence of a gene or the occurrence of another genomic feature. As an illustration of the method, we will focus on occurrences of gene and motifs.

In next Section we will briefly describe what functional motifs are and we will shortly review the empirical evidences that can be found in the existing literature on the interaction between motifs and genes.

The model we propose is a log-linear multivariate intensity Poisson model [15] borrowed from the neuroscience literature (see [16, 17] among others), where these models are quite common, and we show how it can be used to detect positive and negative correlations in a set of words suspected to play a biological function. We apply this method to some motifs in the *E. coli* genome and we show that these models fit the data well.

The rest of the work is organized as follows: in Section III we derive a model of motifs' dependencies starting from a very simple empirical evidence. The model is formally described in Section IV where we also sketch the iterative algorithm used to estimate it. Maximum likeli-

* To whom correspondence should be addressed: d.pirino@sssup.it

hood estimates are reported in Section V while a goodness of fit test is developed in Section VI together with the a-posteriori validation of the hypotheses introduced in Section III. Finally, we discuss our findings in Section VII and we report our conclusions in Section VIII.

II. NON-RANDOM SEQUENCES IN DNA: THE CASE OF FUNCTIONAL MOTIFS.

Functional motifs are short strings of DNA, RNA or even proteins that share the same biological function. Functional motifs differ from structural motifs as they require the aminoacid to be adjacent, while the latter ones require a 3D arrangement. Protein binding sites and *cis* regulatory motifs are typical examples of DNA functional motifs. They are present in all types of genomes from Archaea to humans, both in coding and noncoding regions.

Sometimes a single sequence can perform the same function in a wide range of genomes. In other cases slightly different sequences perform the same function in different species. In such cases a consensus sequence is built that spans a certain number of species [18].

One of the most important open problems in computational genomics nowadays is predicting functional motifs. The most common computational approach to the problem is to compile a list of previously characterized functional motifs and perform a genome-wide scan for overrepresented motifs contained in the list. This approach lies on the assumption that a sequence that is functional in all its occurrences will be more frequent than if it was appearing by chance [19].

Another purely computational approach is just based on retrieving overrepresented words in the genome. As the probability of an over-representation by chance is very low, overrepresented motifs have to be functional at least in some of their occurrences.

The list of motifs that have been identified in past years is quite long, among them there are the gene promoter TATAAT [20, 21], the very frequent uptake signal sequence (USS) AAGTGCGGT present in *H. influenzae* and the USS sequence GCCGTCTGAA of *N. meningitidis* (both analyzed in [22]), the CHI recombinational hotspots GCTGGTGG of *E. coli* [23] and GCGCGTG of *L. lactis* [24]. The latter coincides with the CHI site of *Streptococcus pyogenes*, *Streptococcus pneumoniae*, *Streptococcus agalactiae* and *Streptococcus thermophilus*, as shown in [25]. The same authors find, by predictive modeling, that the motif GAAGCGG is the functional Chi site in the *Staphylococcus aureus*. The prominent role in chromosome replication of the motif GATC in *E. coli* is analyzed in [26]: the replication origin (*oriC*) of the *E. coli* chromosome contains 11 GATC sites in 254 bp, a density that points toward a total rejection of a random accumulation. Moreover GATC-GATC interactions clearly appear when the GATC distribution along the genome is put under investigation. In [27] is shown that, in whole the genome of *E. coli*, GATCNNGATC pairs are underrepresented while the most favored distance between two consecutive GATC occurrences ranges in 1100-1200 bp. On the same line [28] found that a very short distance of $10-20$ nucleotides between GATC motifs is most favorable in SeqA-bound regions of *E. coli*. This suggests at least two different functions for the palindrome GATC. Other motifs operating in bacterial genome are described in [29].

A completely different approach comes from the adoption of chaotic maps for the identification of "non-random" sequences in genomic data (see, among others, [30, 31]). Quite recently, [32] adopted a multifractal spectrum analysis to identify correlations in motif sequences of the human genome. They show that the observed multifractal spectra of all human chromosomes are far away from those expected if the sequences were randomly generated. Notably in [33] it is shown that this spectrum can be surprisingly well fitted, for positive order exponents, with that one of a coupled map lattice.

III. A MODEL OF MOTIFS DEPENDENCIES

We start our analysis from a very simple empirical observation. On the genome of *E. coli* K12 [34] consider the set of motifs GATC, TATAAT, TTGACA, and the CHI recombinational hotspot GCTGGTGG. We further include in our sample the gene position as it is provided from Genbank resources.

We interpret each occurrence (of a motif or of a gene) as an "event" along the genome. If a dependence among occurrences of events exists it must affect the distribution of inter-event distances. Figure 1 reports the observed distribution of the distances between couples of events.

More precisely, the black line with triangles (the Gene→Gene in the legend) is the empirical density of the distances between an occurrence of a gene and the next occurrence of a gene, similarly the red line with stars (the GATC→Gene in the legend) corresponds to the density of the distances between an occurrence of a motif GATC and the next occurrence of a gene, etc. The distances are reported in kilo base-pairs (kbps).

An eye inspection of the empirical densities reveals that there is a notable difference in the structure of the inter-event distances. While the TATAAT→Gene and TTGACA→Gene inter-event distances are peaked around small values, the remaining ones display an empty zone around the origin, especially in the Gene→Gene case, and a shifted and less pronounced peak. Both TATAAT and TTGACA are well-known gene promoters of *E. coli* [35], therefore their corresponding distribution of motif-gene distances is expected to be peaked around small distances. In the other cases it seems that a repulsive effect exists on short distances that avoids a gene to be located near a gene/GATC/CHI locus. Note that while the repulsion between genes can be explained by their finite size (about 1 kbps), motifs are

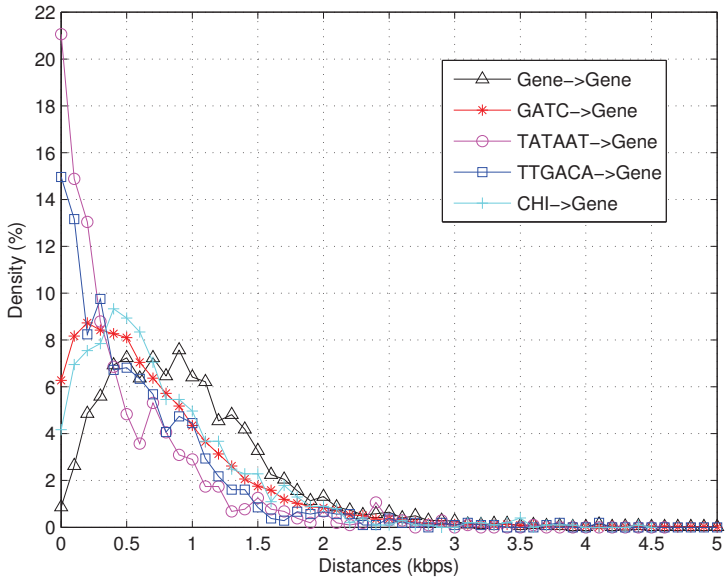

FIG. 1. (Color online) Reports the observed empirical densities (in percentage) of the distances (in kilo base-pairs) between an occurrences of a gene, GATC, TATAAT, TTGACA, and CHI and the subsequent occurrence of a gene. The bin width is set to 100 base-pairs.

practically point-like and therefore, for a random distribution of events, their theoretical density of inter-event distances should be well approximated by a negative exponential, therefore no repulsive effect at short distances is expected.

This simple preliminary analysis points toward the adoption of a model apt to capture attractive as well as repulsive interactions among occurrences of words that are supposed to have a biological function in the genome of *E. coli*. To obtain such a model, we assume that the presence of a motif q in position t in the genome is a random Bernoulli variable, where randomness comes from the stochastic nature of the mutational and evolutionary processes acting on the genome. The probability $P_q(t)$ that the motif q could be found in t depends actually on the local genomic features (other motifs, local GC content, regulatory sequences, genes, non-coding RNA, local chromatin structure...), denoted collectively here as $g(t)$. Since most of these features are unknown or not considered in the analysis, the probability is given by the integral over the distribution of genomic features:

$$P_q(t) = \int d\mu(g(t)) P_q(t|g(t))$$

In this case, the continuum limit of the model would reduce to a (possibly inhomogeneous) Poisson model with parameter $P_q(t) = \lambda_q(t)\, dt$. If we assume that the genome dynamics is approximately invariant under translations over distances much shorter than the size of the genome, then the parameter λ_q does not depend on the location and the distribution of motifs is described by a simple Poisson model.

Now we include interaction among motifs. We denote by $\{n, s\}$ the occurrence of an event of type n in position s and by $N_n(s)$ the cumulative number of events of type n in position s. An additional motif n in position s will modify the probability $P_q(t|g(t))$ of a factor $\delta P_q(t|g(t))/\delta N_n(s) = P_q(t|g(t), \{n, s\}) - P_q(t|g(t))$, where $P_q(t|g(t), \{n, s\})$ is the probability to find motif q at position t conditioned on $g\,(t)$ and on the presence of n in position s. We assume that (i) the effect of each feature on the probability of finding a motif $P_q(t|g(t))$ is small, that is, $|\delta P_q(t|g(t))/\delta N_n(s)| \ll P_q(t|g(t))$; (ii) the genome dynamics is translationally invariant, that is, the interaction between the events $\{n_1, t\}$ and $\{n_2, t + \Delta t\}$ does not depend on the absolute position t but only on the relative position Δt.

Under these assumptions, the effect of a motif of type n in position s on the probability of occurrence of a motif of type q in position t is

$$\begin{aligned}
\frac{\delta P_q(t)}{\delta N_n(s)} &= \int d\mu(g(t)) \frac{\delta P_q(t|g(t))}{\delta N_n(s)} = \\
&= \frac{\int d\mu(g(t)) P_q(t|g(t)) \frac{\delta P_q(t|g(t))/\delta N_n(s)}{P_q(t|g(t))}}{\int d\mu(g(t)) P_q(t|g(t))} P_q(t) = \\
&= \mathrm{E}\left[\frac{\delta P_q(t|g(t))/\delta N_n(s)}{P_q(t|g(t))} \middle| \{q, t\} \right] P_q(t) = \\
&= \mathrm{E}\left[\frac{\delta P_q(0|g(0))/\delta N_n(s-t)}{P_q(0|g(0))} \middle| \{q, 0\} \right] P_q(t)
\end{aligned}$$

where the expectation value is conditioned on the event q in position x using Bayes' theorem and the last step follows from translational invariance [36]. After redefining the quantity

$$K_{q,n}(t-s) \equiv \mathrm{E}\left[\left.\frac{\delta P_q(0|g(0))/\delta N_n(s-t)}{P_q(0|g(0))}\right| \{q,0\}\right]$$

and passing to the continuum limit $P_q(t) = \lambda_q(t)dt$, we obtain for a single event

$$\frac{\delta \log(\lambda_q(t))}{\delta N_n(s)} = \frac{1}{\lambda_q(t)}\frac{\delta \lambda_q(t)}{\delta N_n(y)} = K_{q,n}(t-s) \qquad (1)$$

Now we can consider the joint effect of all events $\{n_j, s_j\}$. Since all the effects are assumed to be small, nonlinear interaction terms among different events can be neglected at first order. Summing the equation (1) over all events, we obtain a Poisson model with parameter

$$\lambda_q(t) \propto \exp\left(\sum_{\text{events } j} K_{q,n_j}(t-s_j)\right)$$

that is, the model reduces to a log-linear intensity Poisson model.

Note that in the DNA there is an inherent asymmetry between the two directions 5'→3' and 3'→5', since the transcription process acts in the 5'→3' direction. This leads to a causality relation in the 5'→3' direction for some sets of motifs or features. In the rest of the analysis we assume that the t axis is 5'→3' oriented and that (iii) an event $\{n, s\}$ has no influence on the occurrence of the event $\{q, t\}$ if $s > t$ (i.e. $K_{q,n}(t-s) = 0$ for $s > t$). This assumption will be relaxed in future work. The reliability of assumptions (i), (ii) and (iii) is confirmed *a posteriori* by a goodness of fit test of our model on a specific dataset (see Section VI) . In particular, note that while hypothesis iii) can be interpreted as an implication of the chemical reading sense of the genome, i) and ii) are additional assumptions that we require to justify the adoption of a log-linear intensity model. Nevertheless they will be discussed in detail in Section VI, where we will give empirical evidence to support them.

IV. FORMAL MODEL

As mentioned above we interpret the occurrence of a motif or of a gene as an event in the tape represented by the genome.

Let $t \in [0, T]$ be the coordinate along a genome, composed by T base-pairs. Let $N_n(t)$ be the counting process of the n-th process. In our specific case $n = 1, 2, 3, 4, 5$ corresponds to the counting process of starting point of a gene, GATC, TATAAT, TTGACA, and CHI occurrences, respectively. In the log-linear intensity model the conditional intensity of having an event of type q at position t is written as

$$\log \lambda_q(t|\mathcal{H}_t) = \mu_q + \sum_{n=1}^{N}\int_0^t K_{q,n}(t-s)\, dN_n(s), \quad q = 1, .., N, \qquad (2)$$

where the *infectious function* $K_{q,n}(t-s)$ describes the effect of an occurrence at time $t-s$ of motif n (trigger) on the instantaneous conditional probability (i.e. $\lambda_q(t|\mathcal{H}_t)$) of having a motif of type q (target) at time t and where the conditioning is given by the filtration [37] generated by all the counting process of the system:

$$\mathcal{H}_t = \sigma(N_n(s) \mid 0 < s < t, n = 1, ...N).$$

In the specification of the model given by equation (2) the quantity $\exp(\mu_q) * dt$ represents the spontaneous probability of having an event q in the infinitesimal portion dt of the genome. The baseline activity μ_q is essential to fit the missing dynamics not captured by the interaction with the other counting processes.

Let $[0, D]$ be the support of the infectious function $K_{q,n}(\tau)$ (i.e. the memory of the system) and let $\Pi = \{t_0 = 0 < t_1 < ... < t_M = D\}$ be a partition of $[0, D]$ with evenly spaced points

$$t_i = i * W, \quad i = 0, .., \frac{D}{W},$$

where we have assumed, without any restriction, that $\frac{D}{W} \equiv M$ is an integer. As in [16] we approximate the infectious function via simple functions [38]:

$$K_{q,n}(\tau) \approx \sum_{k=1}^{M} \alpha_{q,n,k}\mathbb{I}_{[t_{k-1},t_k]}(\tau),$$

where the indicator function $\mathbb{I}_{[t_{k-1},t_k]}$ is defined as:

$$\mathbb{I}_{[t_{k-1},t_k]}(\tau) = \begin{cases} 1 & \tau \in [t_{k-1}, t_k] \\ 0 & \tau \in \mathbb{R}^+ / [t_{k-1}, t_k] \end{cases}.$$

The model (2) is now re-written as:

$$\log \lambda_q(t|\mathcal{H}_t) = \mu_q + \sum_{n=1}^{N}\sum_{k=1}^{M} \alpha_{q,n,k}\, dN_n([t_{k-1}, t_k]), \quad (3)$$

where the random measure $dN_n([t_{k-1}, t_k])$ corresponds to the total number of events of element n in

the interval $[t_{k-1}, t_k]$. In the formalism of [39] a distance $t_k = k\,W$ from an occurrence of process n and the next occurrence of process q is favored anytime that $\alpha_{q,n,k} > 0$, unfavored if $\alpha_{q,n,k} < 0$, and neither favored nor unfavored when $\alpha_{q,n,k} = 0$. Note that the advantage in adopting a log-linear model with respect to a linear one (as in [39]) is that there are no constraints on the parameter space: $\alpha_{q,n,k}$ is allowed to vary in whole real line $\mathbb{R}$.

The log-likelihood of model (2) is given by (see [40])

$$l_q = \int_0^T \log \lambda_q \left(t|\mathcal{H}_t\right)\, dN_q\left(s\right) + \int_0^T \left[1 - \lambda_q\left(t|\mathcal{H}_t\right)\right]\, dt, \tag{4}$$

and is a function of the $U = N \times M + 1$ model parameters $\left(\mu_q, \alpha_{q,n,k}\right)_{q,n=1,...,N;k=1,...,M}$. The total number of parameters of the system is thus $N \times U$. Following [16] and [17] the likelihood (4) can be maximized via the expectation maximization algorithm of [41]. Re-write equation (3) using a reduced index $j = (n-1) * N + k$ obtaining:

$$\log \lambda_q \left(t|\mathcal{H}_t\right) = \sum_{j=0}^{R} \alpha_{q,j}\, I_j\left(t\right),$$

where $R = N \times M + 1$ and $I_{j=(n-1)*N+k}\left(t\right)$ is the number of events of element n in window $[t - k\,W, t - (k-1)\,W]$ and $I_0\left(t\right) = 1$ for all t. In this new notation we have defined $\alpha_{q,0} = \mu_q$. Let $\alpha^0_{q,j}$ be a guess for the model parameters. Define $\gamma^0_{q,j} = \exp\left(\alpha^0_{q,j}\right)$ and apply recursively the iterative algorithm:

$$\gamma^{n+1}_{q,j} = \gamma^n_{q,j} \left[\frac{\sum_{k=0}^{T} I_j\left(k\right)\, \left(N_q\left(k+1\right) - N_q\left(k\right)\right)}{\sum_{k=0}^{T} I_j\left(k\right) \prod_{l=0}^{R} \left(\gamma^n_{q,l}\right)^{I_l(k)}}\right]^{\beta_{q,j}} \tag{5}$$

$$\beta_{q,j} = \frac{\sum_{k=0}^{T} I_j\left(k\right)\, \left(N_q\left(k+1\right) - N_q\left(k\right)\right)}{\sum_{k=0}^{T} I_j\left(k\right) \sum_{l=0}^{R} I_l\left(k\right)\, \left(N_q\left(k+1\right) - N_q\left(k\right)\right)}, \tag{6}$$

where we have imposed an unitary "time-step" (i.e. $dt = 1$). The initial starting point $\alpha^0_{q,j}$ is computed according to the algorithm proposed by [17], which gives a reasonable and easy-to-compute guess of the model parameters. We also implement the stopping rule of [17], that is algorithm (5)-(6) is stopped at the iteration $\bar{n}$ such that [42]:

$$\max_j \left(\frac{\gamma^{\bar{n}+1}_{q,j}}{\gamma^{\bar{n}}_{q,j}} - 1, \frac{\gamma^{\bar{n}}_{q,j}}{\gamma^{\bar{n}+1}_{q,j}} - 1\right) < 10^{-4}, \qquad q = 1, ..., N.$$

The output of the iterative algorithm are the maximum likelihood estimates $\hat{\gamma}_{q,j}$ of parameters $\gamma_{q,j}$ and, as a consequence, of the original parameters: $\hat{\alpha}_{q,j} = \ln\left(\hat{\gamma}_{q,j}\right)$. The rejection of the null hypothesis [43] $\alpha_{q,j} = 0$ is tested using the standard properties of the maximum likelihood estimator [44]. We first compute the t-statistic:

$$t_{q,j} = \frac{|\hat{\alpha}_{q,j}|}{\sigma_{q,j}}, \tag{7}$$

where the standard deviation $\sigma_{q,j}$ is given by:

$$\sigma_{q,j} = \left(\sqrt{-\left[\left(\frac{\partial^2 l_q}{\partial \alpha_{q,l} \partial \alpha_{q,k}}\right)_{l,k=0,...,R}\right]^{-1}}\right)_{j,j},$$

i.e. the diagonal element of the square root of the inverse hessian matrix (changed by sign) of log-likelihood (4). Therefore we reject the hypothesis $\alpha_{q,j} = 0$ with confidence β (or with a p-value [45] $1 - \beta$) if $t_{q,j} \geq \Phi\left(\beta\right)$, where $\Phi\left(\cdot\right)$ is the inverse of the cumulative distribution function of a Gaussian variable with mean zero and unit standard deviation.

Model selection is achieved through AIC criterion. We fix the value of D to 5000 base-pairs [46] and we select the total number M of windows (as a consequence the window width is fixed by $W = D/M$) that minimizes the AIC function:

$$A\left(M\right) = 2\,\left[N \times \left(N \times M + 1\right)\right] - 2 \sum_{q=1}^{N} \hat{l}_q, \tag{8}$$

where $\hat{l}_q$ is the log-likelihood of process q computed in the optimal point $(\hat{\mu}_q, \hat{\alpha}_{q,1}, ..., \hat{\alpha}_{q,N\times M})$. Figure 2 plots the value of $A(M)$ as a function of M. From this plot it is quite clear that the optimal trade-off between the value of the maximized likelihood and the number of model parameters to employ is reached at $M = 25$, according to the AIC criterion. This will be our final choice leading to a total number of $5 \times (5 \times 25 + 1) = 630$ parameters for our system.

An approach similar to that presented in this Section is proposed by [39], however here a linear model is adopted. We believe that a log-linear approach is preferred because naturally incorporates exciting as well as inhibitory connections without any call to constrained maximization procedures. The second main difference between model (2) and the model proposed by [39] lies in the parametrization of the infectious functions. As explained, we adopt a decomposition in simple functions (as suggested in [16]), that, in our view, is preferred to the B-splines adopted by [39] for three main reasons: I) it avoids spurious smoothing of the kernel functions, II) it provides simpler interpretations to each parameters of the model, III) in the B-splines approach the total number of parameters that enter in the model are selected by the AIC criterion, as in our case, nevertheless the choice of the degree of the polynomial of the B-splines remain a little bit arbitrary.

V. MAXIMUM LIKELIHOOD ESTIMATES

In this section, after a brief description of the dataset, we discuss the estimation of the proposed model. We refer to Section VII for the interpretation of our results.

Our dataset is composed by the double strand of the complete genome of *E. coli K12* plus the positions of all the genes, as provided by Genbank. The position of each gene is identified with the position of its first coding base. We find 4490 occurrences of genes, 19120 occurrences of the palindrome GATC, 1036 occurrences of the gene promoter TATAAT, 1057 occurrences of the gene promoter TTGACA and 1008 occurrences of the CHI motif GCTGGTGG.

The estimated baseline activities $\exp(\hat{\mu}_q)$ are reported in Table I. The t-statistics for these parameters are very large (≥ 98) and therefore are omitted. The last column of Table I reports the expected probability of the corresponding event under the uniform model M00 (for more details about models of word occurrences see [47]). This model computes the probability of finding a particular word of length m as $(1/4)^m$, i.e. attributes to each single nucleotide the same probability and does not take into consideration corrections of higher order [48] (for example the abundance of C-G nucleotides, which varies widely both across taxa and within genomes regions).

Maximum likelihood estimates $\hat{\alpha}_{q,n,k}$ of parameters of model (2) are reported in Figure 3. The label on the top horizontal axis indicates the trigger event (process labelled as n in equation 2) while the label on the left vertical axis indicates the target event (process labelled as q in equation 2). We distinguish two levels of significance: parameters with a p-value less then 10^{-5} are marked with a blue cross while parameters with a p-value less then 10^{-6} are marked with a magenta square. A goodness of fit test is achieved through residual analysis and is reported in Section VI. Similarly to what we have done in our preliminary analysis (see Figure 1) we report in Figure 4 the inter-event interval (normalized and in percentage) distribution. More precisely the histogram in sub-figure positioned in the q-th row and n-th column of Figure 4 is the distribution of the distance between event n and q, conditioned on having observed event q (target) after event n (trigger). As in Figure 3 diagonal sub-figures correspond to self-interactions. Each distribution is a normalized histogram with a bin of 1 base-pairs. In each sub-figure we report vertical magenta lines initiated and terminated by a triangle in correspondence of a negative model parameter with p-value less than 10^{-6}. Similarly, we report vertical red lines initiated and terminated by a circle in correspondence of a positive model parameter with p-value less than 10^{-6}. Vertical lines highlight thus particular values in the inter-event interval distribution that have a statistical significance according to log-linear model (2).

	Baseline Probability	
Event	$\exp(\hat{\mu}_q)$	M00
Gene	9.68×10^{-4}	–
GATC	41.34×10^{-4}	$(1/4)^4 = 39.06 \times 10^{-4}$
TATAAT	2.23×10^{-4}	$(1/4)^6 = 2.4414 \times 10^{-4}$
TTGACA	2.28×10^{-4}	$(1/4)^6 = 2.4414 \times 10^{-4}$
CHI	2.17×10^{-4}	$(1/4)^8 = 1.5259 \times 10^{-5}$

TABLE I. Reports (second column) the maximum likelihood estimates of the baseline probability for each process in the system (recall that $dt = 1$ so that $\exp(\hat{\mu}_q)$ coincides with the baseline probability of having a motif q somewhere in the genome). The third column shows the probability of finding a word of the corresponding motif length under the model M00 (see text for details).

VI. GOODNESS OF FIT AND HYPOTHESIS TESTING

Let $\hat{\mu}_q, \hat{\alpha}_{q,n,k}$ be the maximum likelihood estimates of the model parameters defined in equation (3). The model-implied intensities functions $\hat{\lambda}_q(t|\mathcal{H}_t)$, with $q =$

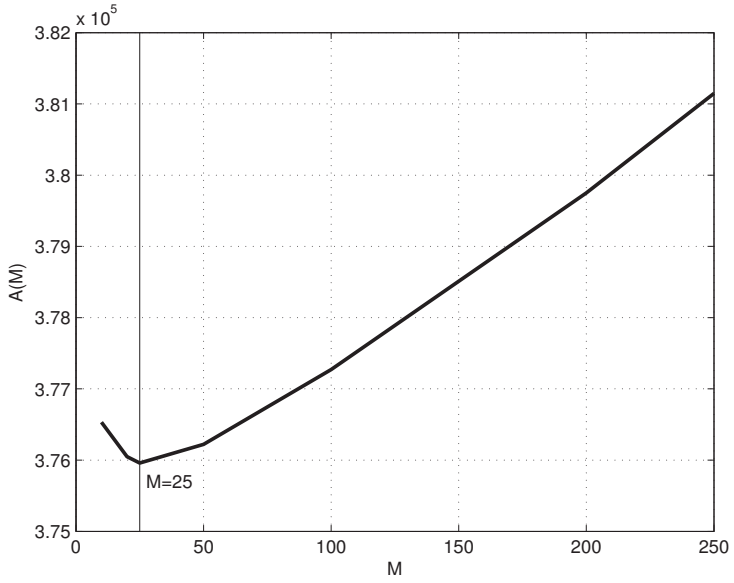

FIG. 2. (Color online) Plots the value of $A(M)$ (see definition 8) as a function of the number M of windows in which the support $[0, D]$ of the infectious function is partitioned. The value for D is set empirically to 5×10^3 base-pairs and the window width W is derived accordingly as $W = D/M$.

$1, ..., N$, are easily computed as:

$$\hat{\lambda}_q\left(t|\mathcal{H}_t\right) = \exp\left[\hat{\mu}_q + \sum_{n=1}^{N}\sum_{k=1}^{M}\hat{\alpha}_{q,n,k}\, dN_n\left([t_{k-1}, t_k]\right)\right]. \tag{9}$$

Note that since $dt = 1$ the previous quantity approximates the probability to have an event q (gene or motif) at position t along the genome. As shown in [49] if u_k^q, with $k = 0, ..., T_q$, where T_q is the total number of events for the q-th element of the ensemble, is a realization of a counting process with conditional intensity $\hat{\lambda}_q\left(t, \mathcal{H}_t\right)$ then the variables:

$$z_k^q = 1 - \exp\left(\int_{u_{k-1}^q}^{u_k^q} \hat{\lambda}_q\left(s, \mathcal{H}_s\right)\, ds\right), \quad k = 1, ..., T_q \tag{10}$$

are uniformly distributed in $[0, 1]$. The order statistics of (10) can be compared with the one of a uniformly distributed variable, i.e. $U_k^q = \frac{k-\frac{1}{2}}{T_q}$. The rationale is that if the model-implied intensity (9) is a correct description of the observed counting process the points

$$\xi_k^q = \left(U_k^q, z_k^q\right), \quad \begin{cases} q = 1, ..., N \\ k = 1, ..., T_q \end{cases} \tag{11}$$

should lie on a 45° line. Figure 5 reports (as thick lines) the observed ξ_k^q for each element of the ensemble considered. The red dotted lines in each plot represent the 99% confidence bands. The model provides a reliable description of the observed counting processes, with the only exception of TATAAT. Nevertheless we have checked the TATAAT fit is quite improved if we use as a promoter TATA instead of TATAAT.

While the goodness of fit test witnesses a general agreement of the data with the model, each single hypothesis introduced in Section III can be tested. As mentioned above we have postponed for further studies the analysis of a non-causal model, thus hypothesis iii) remains just a consequence of the chemical reading sense of the genome. Nevertheless, hypothesis i) can be written as

$$\frac{|\delta P_q(t|g(t))/\delta N_n(s)|}{P_q(t|g(t))} \ll 1,$$

which directly implies that

$$|K_{q,n}(\tau)| \equiv \left| \mathrm{E}\left[\frac{\delta P_q(0|g(0))/\delta N_n(\tau)}{P_q(0|g(0))} \mid \{q,0\} \right] \right| < \mathrm{E}\left[\left| \frac{\delta P_q(0|g(0))/\delta N_n(\tau)}{P_q(0|g(0))} \right| \mid \{q,0\} \right] \ll 1,$$

i.e. the absolute value of the infectious function should be a small quantity, or at least less than one. Apart from the short-range Gene→Gene and CHI→TATAAT interactions, the estimated values of Figure 3 reveals that the interaction is weak and thus that hypothesis i) is valid [50]. Translational invariance of the interaction (hypothesis ii) is confirmed by Figure 6, where a K-fold cross-validation of the empirical densities of Figure 1 (extended to all the pairs of genomic features) is shown. The K-fold cross-validation validation is obtained by slicing the entire sequence of events in non-overlapping slices of 1000 events each, and then producing the histograms of Figure 1 under the random selection of five of these slices. A statistics of these histograms have been obtained over 100 repetitions. Under this condition, if the interaction is translationally invariant then the resulting confidence bands should embrace the distribution obtained using the whole genomic sequence. This is actually what we obtain in Figure 6 (see caption for more details), in particular for features with a higher populated statistics (Gene and GATC).

VII. DISCUSSION

The estimates reported in Table I suggest that, apart from the CHI case, the baseline probability of a motif event is in line with that of the random model M00. The CHI case presents, on the contrary, a level approximately one order of magnitude larger than what could be expected from a uniformly random draw.

An inspection of the estimated infectious function in Figure 3 confirms the intuition suggested by Figure 1. The CHI motif has a negative (and highly significant) correlation on the probability of a gene event (first row and last column of Figure 3). Thus the empty zone in the CHI→Gene distribution (first row and last column of Figure 4) is explained by this repulsive effect. A similar evidence is found for the GATC motif, which presents a negative correlation with gene occurrences for distances approximately less than 0.4 kbps and a positive correlation nearly at 2.2 kbps. The esamers TATAAT and TTGACA show, as it should be for gene promoters, a positive short-run correlation with gene occurrences and this explains the peak positioned around small distances in the distribution of TATAAT→Gene and TTGACA→Gene interval distributions (first row and third/fourth columns of Figure 4).

The repulsive effect of CHI and GATC on gene occurrences can be easily explained by their role in genome maintenance. In particular, the CHI motif plays a role during DNA strand breaking and repair. A coding sequence close to the motif has an higher probability of being spoiled during the repair process, therefore CHI motifs near genes are negatively selected.

It is interesting to note that the repulsive force of CHI and GATC against a gene occurrence is associated to the inhibition of the TATAAT and TTGACA promoters. In fact the occurrence of TATAAT is inhibited by both GATC and CHI at short distances (third row and second/fifth columns of Figures 3-4), while TTGACA is inhibited solely by CHI (fourth row and fifth column of Figures 3-4). This interaction is roughly symmetric: an occurrence of TATAAT and TTGACA reduces the probability of a GATC occurrence (second row and third/fourth columns of Figures 3-4).

Here we notice that the proposed model is capable to detect both direct and indirect interactions, nevertheless, it cannot discriminate among them.

Moving on to self-interactions, the correlation between two gene occurrences is highly negative in the range from 0 to 1 kbps and thus becomes positive in the range $1.2 - 1.6$ kbps. This profile is mainly explained by the mean length of a gene occurrence, which is approximately 1 kbps and therefore prohibits a new gene event in this range. The remaining self-interactions are significantly different from zero (and positive) only for GATC and TATAAT. The GATC→GATC infectious function reveals a positive feedback for GATC occurrences at short distances and at approximately 1 kbps, a result in line with the findings of [27]. Finally, the TATAAT→TATAAT case shows a quite persistent and positive self-interaction.

VIII. CONCLUSIONS

The analysis presented in this paper starts from a very simple empirical evidence: the distance between a motif and the gene start codon (i.e. the word ATG) depends strongly on the motif. This fact suggests the presence of correlation of different type between motifs and gene occurrences. We show that such a dependence exists not only between a motif and the start codon but also among different motifs regulating the expression of the same gene. This empirical finding suggests the adoption of a multivariate model apt to capture positive as well as negative correlations among events in the genome under study, where an event is defined as an occurrence of a particular DNA motif.

In particular, we have shown that a multivariate Poisson process with log-linear intensities is capable to catch these features. This result confirms, together with previous studies [3–14], the ductility of this class of processes in describing of wide range of physical as well as biological phenomena.

The main objection that can be moved against our model hinges on causality. In fact the model, as largely explained in the Introduction, is originally designed for photon counting, cavities with feedback, neural interactions, earthquake aftershocks and other phenomena, and therefore is a causal model. Nevertheless in model (2) the variable t is a position variable and therefore the model is not required to be causal, but rather locally dependent. However, as mentioned in Section III, there is a preferred direction along the genome coordinate, and the introduction of a non-causal model could result in a reduction of simplicity without any fundamental improvement. In this paper for simplicity we assumed a causal sense 5'→3' for the genome dynamics. The extension to a non-causal framework would be a feasible development and it is postponed for future research.

Our analysis confirms the role of TATAAT and TTGACA as gene promoters. Most notably, we confirm the prominent role in genome maintenance of the CHI and GATC motifs of *E. coli*, which are well-known to be involved in DNA repair and replication [29]. In fact we show that a negative correlation exists between an occurrence of CHI or GATC and the subsequent occurrence of a gene (and of a gene promoter), a feature essential in preserving, during genome repair or replication, the information contained in genes.

Finally, our analysis of the goodness of fit shows that the proposed model is a good description of the process and it could be useful for more detailed studies of the existing interactions between motifs and genomic features.

ACKNOWLEDGMENTS

L.F. acknowledges support from CSIC (Spain) under the JAE-doc program. Work funded by grants AGL2010-14822 (MICINN, Spain) to M. Pérez-Enciso, and Consolider project (MICINN, Spain) to the Centre of Research in Agricultural Genomics.

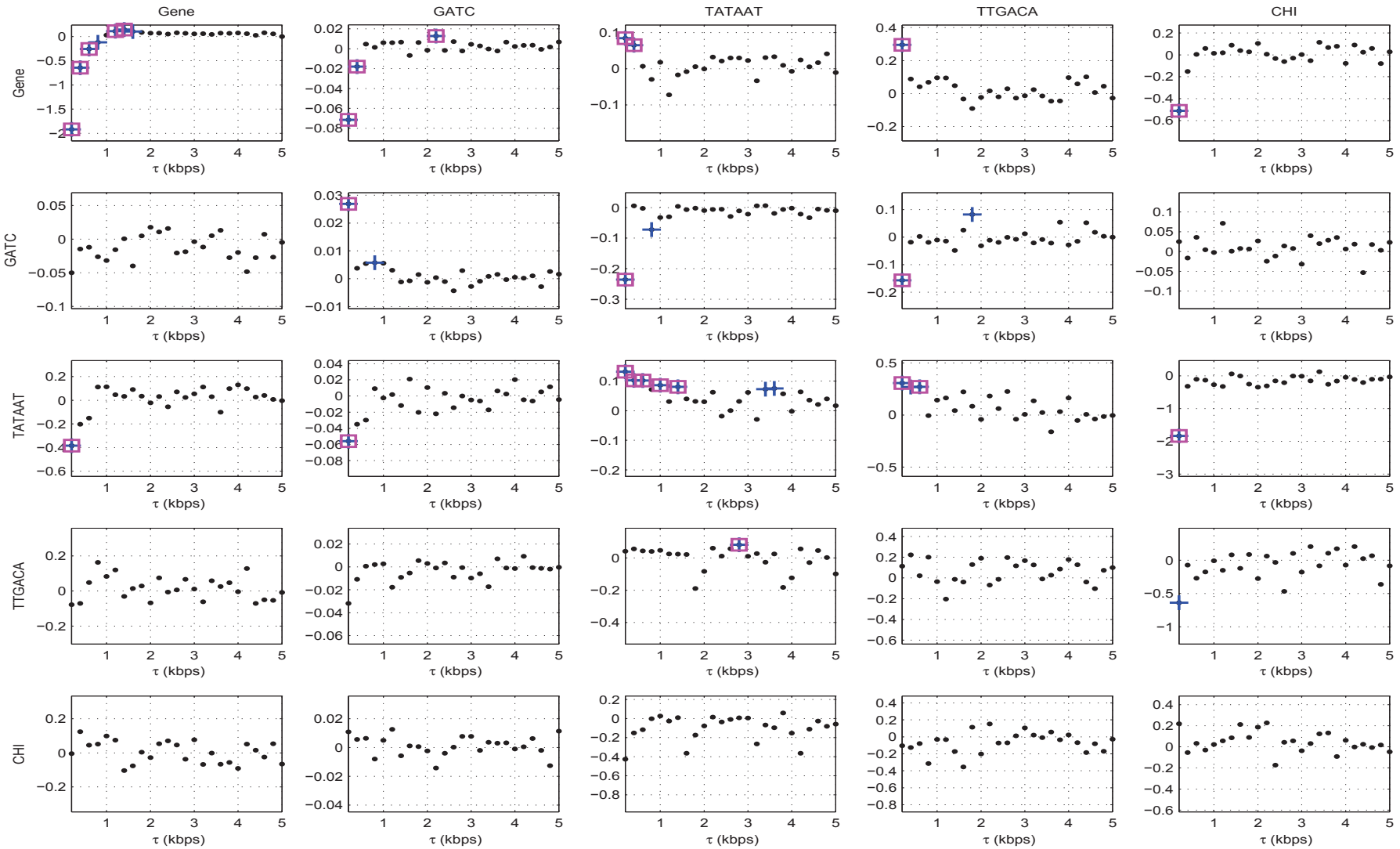

FIG. 3. (Color online) Estimated infectious function $K_{q,n}(\tau)$ (dark points) for the system composed by all genes and the motifs GATC, TATAAT, TTGACA and CHI. The labels on the top horizontal axis indicate trigger events (i.e. the index n in $K_{q,n}(\tau)$) while the ones on the vertical axis indicate target events (i.e. the index q in $K_{q,n}(\tau)$). The abscissa reports the distance τ between events in kilo base-pairs. Blue crosses indicate p-values less than 10^{-5} while magenta squares mark parameters with p-value less than 10^{-6}.

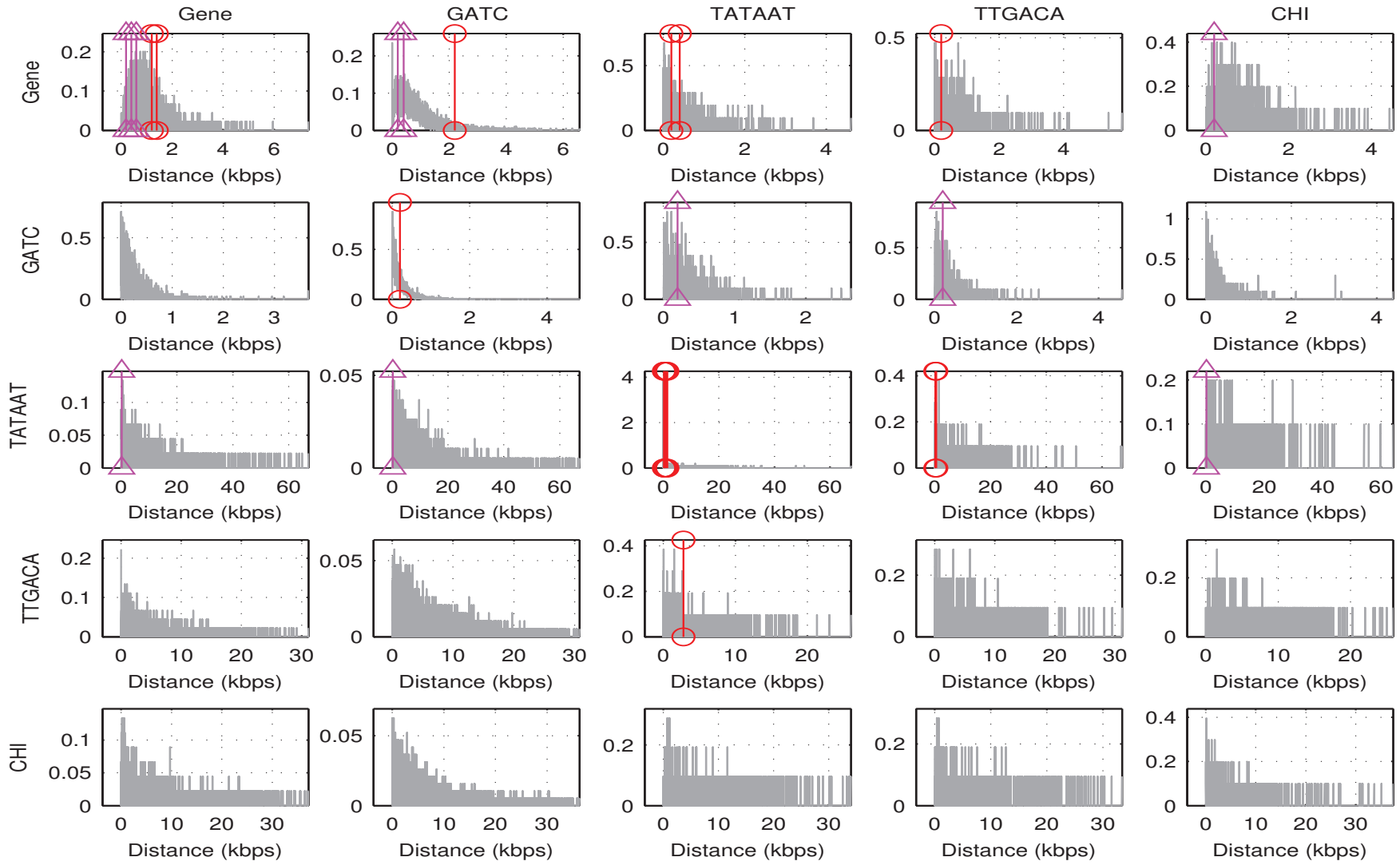

FIG. 4. (Color online) Inter-event distances distribution (in percentage) for the system composed by all genes and the motifs GATC, TATAAT, TTGACA and CHI. The labels on the top horizontal axis indicate trigger events while the ones on the vertical axis indicate target events (see caption of Figure 3 for more explanations). Vertical magenta lines initiated and terminated by a triangle highlight event distances that correspond to negative parameters of model (2) with a p-value less than 10^{-6} while vertical red lines initiated and terminated by a circle highlight event distances that correspond to positive parameters of model (2) with a p-value less than 10^{-6}. The spatial coordinate on the abscissa of each sub-plot is expressed in kilo base-pairs and the bin width is set to a single base-pairs.

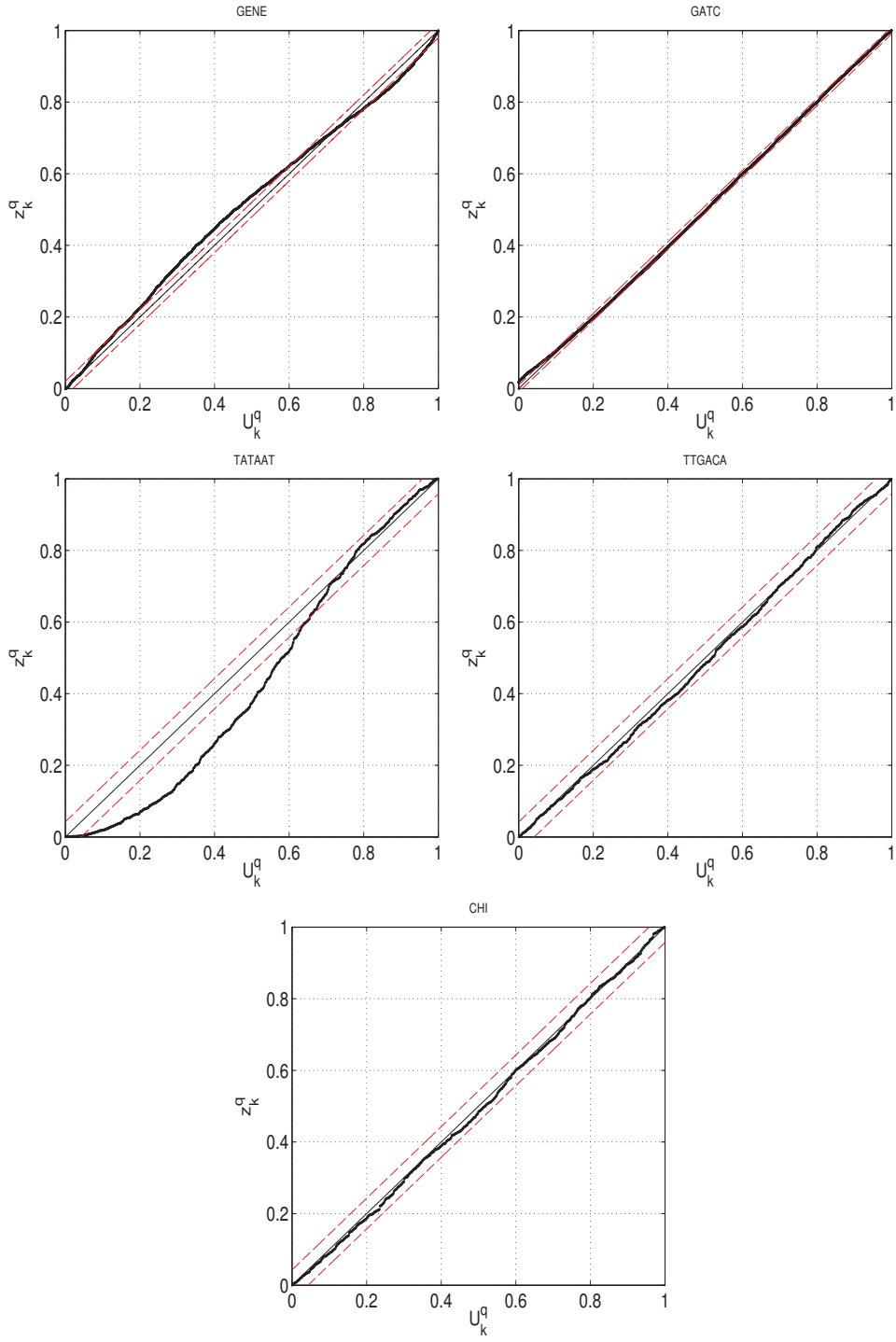

FIG. 5. (Color online) Reports (as thick black line) the points $\xi_k^q = (U_k^q, z_k^q)$ defined in equation (11), where k indexes the events of the q-th element of the ensemble composed by Gene, GATC, TATAAT, TTGACA, and CHI occurrences. Each sub-plot corresponds to a different element of the ens-amble as reported in the corresponding title. Red dotted lines represent 99% confidence bands, while the thin black line represents the 45° line.

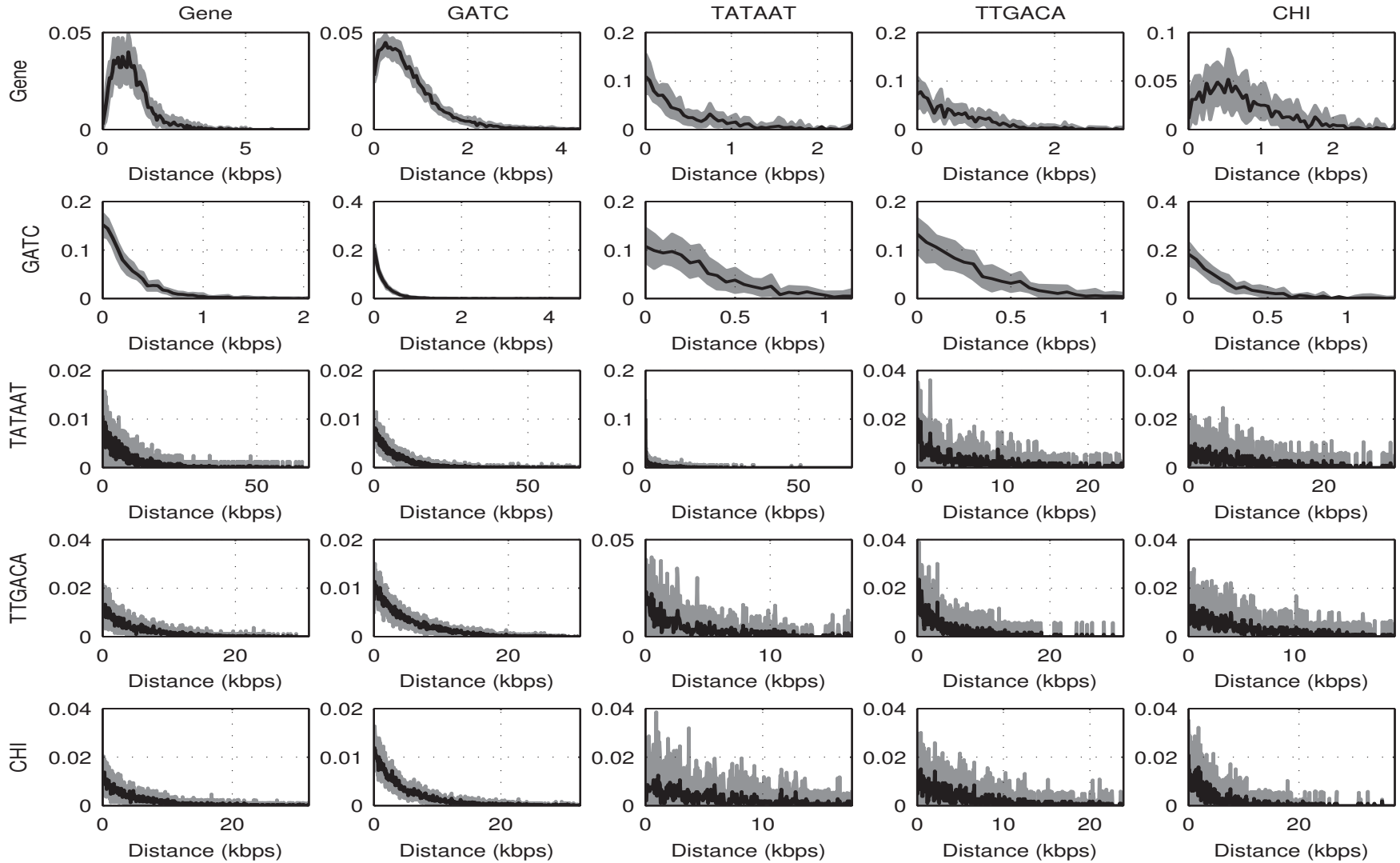

FIG. 6. (Color online) K-fold cross validation of the inter-event interval distribution (empirical densities here are not in percentage for practical reasons, while the bin width is the same of Figure 1) for the system composed by all genes and the motifs GATC, TATAAT, TTGACA and CHI. The black lines plot the distribution computed with the complete sequence of events. The gray shaded area is the area between 5%-95% confidence bands computed with 100 random slicings of the sequence of events in five non-overlapping and non-consecutive slices of 1000 events. The spatial coordinate on the abscissa of each sub-plot is expressed in kilo base-pairs.

[1] For example, if the events under study are earthquake shocks, S_t can be a measure of their magnitude.
[2] A. G. Hawkes, Biometrika **58**, 83 (1971).
[3] S. Ohsuka, Y. Ogata, and Y. Tamura, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **384**, 477 (1997).
[4] H. M. Wiseman and G. J. Milburn, Physical Review A **46**, 2853 (1992).
[5] B. E. A. Saleh and M. C. Teich, Physical Review Letters **58**, 2656 (1987).
[6] M. C. Teich and B. E. A. Saleh, Quantum Optics **1**, 153 (1989).
[7] B. E. A. Saleh and M. C. Teich, Proceedings of the IEEE **80**, 451 (1992).
[8] T. Pechácek and V. Karas, Proceedings of RAGtime 8/9: Workshops on black holes and neutron stars (2007).
[9] T. Pechácek, V. Karas, and B. Czerny, Astronomy and Astrophysics **487**, 815 (2008).
[10] Y. Ogata, Pure and Applied Geophysics **155**, 451 (1999).
[11] A. Helmstetter and D. Sornette, Journal of Geophysical Research **108** (2003).
[12] J. Zhuang and Y. Ogata, Physical Review E **73** (2006).
[13] D. Sornette, S. Utkin, and A. Saichev, Physical Review E **77** (2008).
[14] F. Deschâtres and D. Sornette, Physical Review E **72** (2005).
[15] Here for *multivariate intensity Poisson model* we mean a set of Poisson processes whose intensities (i.e. instantaneous probabilities of an occurrence) may interact each other.
[16] E. S. Chornoboy, L. P. Schramm, and A. F. Karr, Biological Cybernetics **59**, 265 (1988).
[17] M. Okatan, M. A. Wilson, and E. N. Brown, Neural Computation **17**, 1927 (2005).
[18] K. D. MacIsaac and E. Fraenkel, PLoS Comput Biol **2**, e36 (2006).
[19] M. Frith, Y. Fu, L. Yu, J. Chen, U. Hansen, and Z. Weng, Nucleic acids research **32**, 1372 (2004).
[20] D. Pribnow, Proceeding of the National Academy of Sciences **72**, 784 (1975).
[21] M. S. Fenton and D. G. Jay, Proceeding of the National Academy of Sciences **98**, 9020 (2001).
[22] H. O. Smith, M. L. Gwinn, and S. L. Salzberg, Research in Microbiology **150**, 603 (1999).
[23] G. R. Smith, S. M. Kunes, D. W. Schultz, A. Taylor, and K. L. Triman, Cell **24**, 429 (1981).
[24] I. Biswas, E. Maguin, S. D. Ehrlich, and A. Gruss, Proceeding of the National Academy of Sciences **92**, 2244 (1995).
[25] D. Halpern, H. Chiapello, S. Schbath, S. Robin, C. Hennequet-Antier, A. Gruss, and M. El Karoui, PLoS Genetics **3**, 1614 (2007).
[26] W. Didier and J. Casadeus, Nature Reviews Microbiology **4**, 183 (2006).
[27] A. Hénaut, T. Rouxel1, A. Gleizes, I. Moszer, and A. Danchin, Journal of Molecular Biology **257**, 574 (1996).
[28] M. A. Sánchez-Romero, S. J. W. Busby, N. P. Dyer, S. Ott, A. D. Millard, and D. C. Graingerc, mBio **1** (2011).
[29] F. Touzain, M. A. Petit, S. Schbath, and M. El Karoui, Nature Reviews Microbiology **9**, 15 (2011).
[30] Z. Yu and K. Anh, V.and Lau, Physical Review E **64** (2001).
[31] Z. Yu and K. Anh, V.and Lau, Physical Review E **68** (2003).
[32] A. Provata and P. Katsaloulis, Physical Review E **81** (2010).
[33] A. Provata and C. Beck, Physical Review E **83** (2011).
[34] The complete genome sequence can be downloaded from Genbank using accession number $U00096$ and is composed by 4639810 base-pairs.
[35] S. S. Singh, A. Typas, R. Hengge, and D. C. Grainger, Nucleic Acids Research **30**, 1 (2011).
[36] The assumption of small effects ensures that the ratio $\frac{\delta P_q(t|g(t))/\delta N_n(s)}{P_q(t|g(t))}$ is small in absolute value and therefore cannot diverge.
[37] The filtration is a mathematical artifact to model the accumulation of information from the past.
[38] Any real function can be approximated in this way.
[39] G. Gusto and S. Schbath, Statistical Applications in Genetics and Molecular Biology **4**, Article 24 (2005).
[40] A. F. Karr, Annals of Statistics **15**, 473 (1987).
[41] A. P. Dempster, N. M. Laird, and D. B. Rubin, Journal of the Royal Statistical Society **39**, 1 (1977).
[42] The tolerance in [17] is set to 10^{-5}, nevertheless we have found that a tolerance of 10^{-4} gives good estimate and saves time.
[43] Null hypothesis is referred here as the benchmark situation in which correlations among events are absent.
[44] R. Davidson and J. G. MacKinnon, *Econometric Theory and Methods* (Oxford University Press, 2004).
[45] The p-value is the probability to observe a value of the statistic defined in equation (7) at least as extreme as the one we observe in the data, having assumed the null-hypothesis of no-correlations. The smaller this probability the higher the statistical significance of the results. In what follows we will show that we can find p-values as small as 10^{-6}.
[46] We do not find very informative to include larger correlations. Increasing D has as main effect to slow down convergence.
[47] S. Robin, S. Schbath, and V. Vandewalle, BMC Bioinformatics **8**, 1 (2007).
[48] Note that for the case of genes' occurrences this probability varies from gene to gene, having each of them a different length.
[49] E. N. Brown, R. Barbieri, V. Ventura, R. E. Kass, and L. M. Frank, Neural Computation **14**, 325 (2002).
[50] Note that strong inhibitory Gene→Gene short-range interactions are expected simply because genes occupy a non-negligible length and cannot overlap.